DOCUMENT RESUME

ED 110 048 IR 002 327

AUTHOR Porch, Ann
TITLE An Analysis of Methods for Preparing a Large Natural Language Data Base.
INSTITUTION Southwest Regional Laboratory for Educational Research and Development, Los Alamitos, Calif.
REPORT NO SWRL-TM-5-71-02
PUB DATE 16 Feb 71
NOTE 29p.

EDRS PRICE MF-$0.76 HC-$1.95 PLUS POSTAGE
DESCRIPTORS Computers; *Cost Effectiveness; *Data Bases; Data Processing; Electronic Data Processing; *Equipment; *Information Processing; Information Storage; *Input Output Devices; Man Machine Systems; Office Machines; On Line Systems; Optical Scanners; Typewriting
IDENTIFIERS Administrative Terminal System; ATS; Cathode Ray Tube Terminals; CRT; Dataplex; Flexowriter; Keypunches; Magnetic Tape Selectric Typewriter; MTST; Optical Character Scanning; Teletypes

ABSTRACT

Relative cost and effectiveness of techniques for preparing a computer compatible data base consisting of approximately one million words of natural language are outlined. Considered are dollar cost, ease of editing, and time consumption. Facility for insertion of identifying information within the text, and updating of a text by merging with another text are given special attention. It is concluded that Magnetic Tape Selectric Typewriter (MTST) and Telterm2 (a cathode ray rube terminal) are two highly effective methods of text preparation. The decision of which to use on a particular project would depend on available funds and possible peripheral uses for the equipment. Criteria for making such a decision are discussed. (Author)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED110048

SWRL

SOUTHWEST REGIONAL LABORATORY
TECHNICAL MEMORANDUM

DATE: February 16, 1971

NO: TM 5-71-02

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

TITLE: AN ANALYSIS OF METHODS FOR PREPARING A LARGE NATURAL LANGUAGE DATA BASE

AUTHOR: Ann Porch

ABSTRACT

Relative cost and effectiveness of techniques for preparing a computer-compatible data base consisting of approximately one million words of natural language are outlined. Considered are dollar cost, ease of editing, and time consumption. Facility for insertion of identifying information within the text, and updating of a text by merging with another text are given special attention. It is concluded that MTST and Telterm2 are two highly effective methods of text preparation. The decision of which to use on a particular project would depend on available funds and possible peripheral uses for the equipment. Criteria for making such a decision are discussed.

IR 000 327

This document is intended for internal staff distribution and use. Permission to reprint or quote from this material in any document, wholly or in part, should be obtained from SWRL, 11300 La Cienega Boulevard, Inglewood, California, 90304.

AN ANALYSIS OF METHODS FOR PREPARING
A LARGE NATURAL LANGUAGE DATA BASE

I. Introduction

The speed and versatility of the computer for processing natural language text is an important aspect of current research. The proliferation of written information has given birth to numerous computerized information retrieval systems which must cope with natural language. Linguists, educators, social scientists and scholars in the humanities are closely scrutinizing language usage in order to develop models and systems in their respective fields. As the computer industry develops more sophisticated approaches with a "conversational mode" between the user and the computer, the processing of the natural language of human beings takes on new importance. With the increase in special purpose programming languages, compiler design and development, applications increase and more natural language techniques develop as a by-product.

A number of problems arise in natural language or "text" processing by a computer, many of which occur long before the data reaches the computer. Natural language comes in many forms, none of which are directly computer compatible. Spoken language can be recorded on magnetic tape or printed in a book or magazine, but computers are not yet generally equipped to take these forms as direct input to their complex electronic circuitry. Computers need information in the form of on-off codes, commonly referred to as "bits" and represented as 1's and 0's in groups of 6 or 8 called "bytes," with each such group usually representing a single character. For example, the letter A

may be represented by 1000001[1] or 11000001[2] or 110001[3], depending on the computer used to process the data. Generally speaking, computer compatible information is stored in one of three basic forms: punched cards, punched paper tape, or magnetic tape. The problem arises in finding an accurate, easy, and cost-effective method of transforming voice recordings or printed text into a form directly accessible by the computer.

II. Overview of Systems

In an attempt to discover an effective method for preparing text acceptable to the computer, a comparative study was undertaken of devices which might be used to prepare a one million word file of natural language text. The systems analyzed were KEYPUNCH, TELETYPE, FLEXOWRITER, MTST, DATAPLEX, ATS, TELTERM2 (CRT), and OPTICAL CHARACTER SCANNING.

Investigation was limited to techniques immediately available in the greater Los Angeles area, and all prices quoted are those applicable to a non-profit research or educational institution. In order to maintain directly parallel prices for easy comparison, any in-house facilities (such as a Keypunch machine or computer) accessible to the researcher were not considered cost free. In each case, commercial prices are given.

An overview of each system was made by contacting vendors and reading the printed material available on each system. Items considered were cost in dollars, time consumption in operation, and ease of editing.

---

[1] ASCII (American Standard Code for Information Interchange)

[2] EBCDIC (Extended Binary Coded Decimal Interchange Code)

[3] BCD (Binary Coded Decimal)

All cost estimates were made on the basis of one million words of text. All terminal rentals are considered on a one-year lease basis only, total rental figures are for one full year. Cartridge, cassette, and disk storage costs are given on the basis of a full 5,000,000 characters being stored at one time, which could be avoided in practice by more frequent dumping to magnetic tape if a sufficiently flexible magnetic tape updating facility is available. Magnetic tape purchase cost was not considered, since it would remain constant regardless of the system employed.

Keypunch (IBM 029)

The keypunch is an off-line method for storing information on lightweight cardboard by mechanically punching holes in the cards. The arrangement of the holes represents coded data. A high noise level is created by the card punching. Keying may be done on an IBM 029 keypunch machine and hard copy may be obtained by "listing" the cards using a computer. A trained keypunch operator is required. Keypunch training takes an average of six weeks. Typing errors made and caught during entry are correctable by using a verifier (which resembles a keypunch) and duplicating the card up to the error and then typing in the correction. Edit operations such as insertions, deletions, etc. are difficult since many cards may have to be repunched to take care of the corrections. The best approach for insertion is to leave a number of blanks at the end of the card, so that the card may be duplicated up to the needed insertion and the insertion made without overflowing into another card. The rest of the card must be retyped,

since the old card is moved out of position for further duplication by the typing of the insertion. Programs written to process cards produced in such a manner must be written (or available) to treat multiple blanks.

Problems existing with storage on cards include such things as wrapping of cards stored in non-standard humidity or temperature. The bulk of the 62,500 cards[4] required to hold one million words is enormous, and reordering a deck of dropped cards can be difficult if they are not sequence numbered. For these reasons, and others, it is convenient to convert the cards to magnetic tape. Such conversion can be done by using card to tape equipment.

Table 1
KEYPUNCH[5]

Item	Breakdown	Total
Keypunch purchase	$3490 + $28.75/mo maintenance	$3835 ($3490 + $345/yr. maintenance)
Keypunch rental	029-$77/mo.	$924 (12 mo. @ $77 ea.)
Storage cost (card)	$1.15/1,000 + $12 set up	$92.50 (70,000 cards)
Conversion to magnetic tape	$1.00/2,500 cards	$28 (70,000 cards)
Total per year if leasing		$1044.50
Total if purchasing		$3955.50

[4] Each card may have up to 80 characters. Since the million word data base would contain approximately 5,000,000 chracters, 62,500 cards (5,000,000 ÷ 80) would be required.

[5] Prices quoted by: Mr. Bob Wright, IBM, 445 S. Figuroa, Los Angeles, (213) 620-1830 (IBM 029 Keypunch, cards); Mr. Harold Hackman, SBC, 11300 La Cienega Boulevard, Inglewood, California 90304 (213) 776-5900 (card-tape conversion)

Teletype (Western Union Terminal 33)

The Teletype is an off-line method for storing information on paper tape by mechanically punching holes in the tape with eight punched positions per code. Keying may be done on a standard teletype keyboard with simultaneous production of paper tape and hard copy.

Typing errors made and caught during entry can be corrected by the use of "control" characters which indicate to the computer that either the character immediately preceding or the present line should be ignored when the paper tape is processed on-line. These codes can be followed by the corrected copy. Edit operations such as character insertions can theoretically be accomplished by duplicating the paper tape, stopping it exactly at the appropriate spot, adding the insertion and continuing the duplication. In practice, such accuracy in positioning the paper tape is virtually impossible as in the case of punch cards. Although paper tape itself is directly computer compatible, it is bulkier than magnetic computer tape and more susceptible to problems created by handling. Approximately 55,000 feet[6] of paper tape would be required to store the full one million words. Conversion from paper tape to magnetic tape can be done directly by computer.

---

[6] Each inch of paper tape contains 8 codes; one foot contains 96 codes. Therefore, 5,000,000 characters would take 52,083.3 feet (5,000,000 ÷ 96). A paper tape is 1000 feet long and contains 96 codes per foot. Thus, 19,200 five character words could fit on one roll if no errors were made which required extra codes. For purposes of estimating, the figure 20,000 words per roll was used. It is significant to note that the full million word data base would take about 55,000 feet of paper tape, which would be cumbersome and fragile to handle. The estimate on conversion to magnetic tape is purposely high. The time would depend on operator and machine efficiency.

Table 2
TELETYPE[7]

Item	Breakdown	Total
Terminal purchase	$877 + $100 (complex) + $50 (to wire the interface adapter)	$1027
Terminal rental	$55/mo 25/mo for data phone	$960 (12 mo. @ $80 ea.)
Paper tape	$1.15/1000 ft. roll	$63.25 (55 rolls @ $1.15 ea.)
Conversion to magnetic tape	$25/hr. of computer time	$250 (10 hrs. @ $25 ea.)
Total if leasing		$1273.25
Total if purchasing		$1340.25

Flexowriter 2301 (Friden)

The Flexowriter is an off-line system for storing information on punched paper tape very much like teletype. Keying is done on an electric typewriter keyboard with paper tape punch and simultaneous production of hard copy and paper tape with eight punch positions per code. Typing errors made and caught during entry are correctable by moving the paper tape back and blocking out the error with the tape feed key. The correct code may then be entered. After-the-fact correction such as insertions, deletions, etc. can be accomplished by

[7] Prices quoted by Mr. Jacobson, Teletype Corp., 5720 E. Washington Boulevard, Los Angeles, (213) 724-6040. (Exchange data terminal 33 with paper tape reader/punch and coupler); Mr. Bob Bewak, Stat Tab Data Service Center, 1519 Olympic Boulevard, Los Angeles, 90015, (213) 381-7251.

playing the punch tape back at 145 words per minute. A new paper tape is punched simultaneously with playback, and updating can be done by stopping the original tape and making the correction. Positioning the tape at the appropriate place is difficult, but more easily accomplished than with teletype. The original-to-edit process may be repeated as often as necessary. Although the paper tape produced by the Flexowriter is directly machine compatible, the placement of the sprocket holes necessitates the reversing of the tape for read-in to a computer. Read-in under such conditions causes all lower case to be capitalized and causes problems with numeric and punctuation codes. As with teletype, approximately 50,000 feet of paper tape would be needed for data storage.

Table 3
FLEXOWRITER[8]

Item	Breakdown	Total
Terminal purchase		$3300
Terminal rental	$105/mo.	$1260 (rms @ 105 ea.)
Paper tape cost (see comment #2)	$1.15/roll (need 50)	$57.50 (50 rolls @ $1.15)
Conversion (see comment #2)	$25/hr. of computer time	$250 (10 hr. @ $25 ea.)
Total of lease		$1567.50
Total of purchase		$3607.50

[8] Prices quoted by Mr. Michael A. Jackson, Singer-Friden Division, 1720 Beverly Boulevard, Los Angeles 90026, (213) 483-4800 (Flexowriter 2301)

MTST (IBM Magnetic Tape Selectric Typewriter)

The MTST is an off-line system for storing information on magnetic cartridges which have a capacity of approximately 24,000 characters. Keyboarding is done on an IBM Selectric Typewriter with simultaneous production of cartridge and hard copy. A trained MTST operator is required for keyboarding. Initial training consists of 2-4 full days in a special course provided free by IBM. Typing errors made and caught during entry are correctable merely by backspacing and retyping. For editing, the tape cartridge can be played back at 150 words per minute, with such changes to content as insertions, deletions, and substitutions accomplished by recording onto a second tape the material as it is played back from the original. The corrections can be typed by stopping the play back at appropriate points. If "reference codes" are inserted at intervals in the original, the search scan can run through the tape at 10,000 words per minute, stopping at the position after the specified reference code for insertions at that point. The original-to-edit process may be repeated as often as necessary. The final step is transfer from MTST cartridge directly to magnetic tape. If further editing is required, the magnetic tape can be converted back to MTST cartridges for use on the MTST device. Approximately 210 cartridges would be needed for data storage.[9]

---

[9] Each cartridge holds 24,000 characters. Therefore, 5,000,000 characters would take 208.3 cartridges (5,000,000 ÷ 24,000). The $3150 cartridge cost would probably be lower in actual practice, since cartridges could be used several times, thus lowering the total number needed.

Table 4
MTST[10]

Item	Breakdown	Total
Terminal purchase		$10,035
Terminal rental	$257/mo.	$3084 (12 mo @ $257 ea.)
Cartridge cost	$15 ea. (need 210 @ 24,000 ch. ea.)	$3150 (210 @ $15 ea.)
Conversion	$2.50/cartridge	$355 (210 @ $250 ea.)
Total if leasing		$6589
Total if purchasing		$13,540

Dataplex (Data Instruments)

The "Dataplex" is an off-line keying, on-line editing system, using an IBM Selectric typewriter recording simultaneously on paper and on a cassette which holds up to 50,000 characters. The system is designed to work with its own 12k computer, utilizing software created specifically for a particular job. Typing errors made and caught during entry are correctable merely by backspacing and retyping. Errors caught before the end of the cassette can be corrected by typing correction commands. There is no playback capability at this stage. Further editing can be done in batch mode either by hand carrying, mailing or tele-processing the original cassette and a corrections cassette to the processor which

[10] Prices quoted by Mr. Chuck Zander, IBM, 9045 Lincoln Boulevard, Los Angeles, (213) 670-8350 (MTST with code conversion and reverse search, cartridges in lots of 150 or more); Mr. Del Seraphine, Autographics, 751 Monterey Pass Road, Monterey Park 91754, (213) 263-2184.

converts to magnetic tape. The processor reads in at a rate of 200 characters per second. The processing could be done by Data Instruments on a service bureau basis, utilizing a program written by the user or specially prepared by Data Instruments Staff. The update editing of the magnetic tape can be carried out any number of times by submitting further correction cassettes. The complexity of the commands required for the correction cassette would be determined by the sophistication of the software produced. Magnetic tape is always in most updated form and hard copy is produced by line printer at each step after the first. Approximately 100 cassettes would be required for data storage.[11]

Table 5
DATAPLEX[12]

Item	Breakdown	Total
Terminal purchase		$4700
Terminal rental	$104/mo.	$1248 (12 mo. @ $104 ea.)
Cassette cost	$5 ea. (need 100 @ 50,000 ch. ea.)	$500 (100 @ $5 ea.)
Updating processing	$24/hr. (need 42 hr. per editing pass)	$3024 (126 hr. @ $24 ea.)
Total if leasing		$4772
Total if purchasing		$8224

[11] Each cassette holds 50,000 characters. Therefore, 5,000,000 characters would take 100 cassettes (5,000,000 ÷ 50,000).

[12] Prices quoted by Mr. Ed Russo, Data Instruments Company, 16611 Roscoe Place, Sepulveda, California 91343, (213) 893-6644. With possible exception of cassette cost, costs stated are not flexible. In addition, approximately one week minimum of programmer time must be invested.

TELTERM2 (Delta Data Systems)

The Telterm2 is a keyboard CRT terminal with up to 3,000 characters of built-in memory. The terminal can also be used on-line, connected by phone lines to a computer. Data can be transferred off-line onto a cassette tape attachment for later further editing. Each cassette can hold up to 50,000 characters. The CRT screen displays 27 lines of 80 characters each, and immediate roll-up and roll-down capability is available to display other portions of the full 3,000 character memory. Since only the data up to a carriage return is contained in memory, and not full 80 character line, up to 100-150 lines of text may be readily available for editing and direct display. Errors made and caught during entry are correctable by backspacing and retyping. Errors caught before transmission of the full 3,000 characters of memory to the cassette are correctable by moving the cursor to the appropriate position and retyping. Also available are insert and delete function keys. To insert new material, the user merely positions the cursor at the point where the insertion is to be made, and pushes the INSERT key. He may then type in whatever insertion he wants, and the Telterm2 will automatically "wrap around" the rest of the data contained in its memory. An END INSERT key is used to complete the insertion. A similar procedure produces deletions of either single characters or complete lines. A format mode of data entry allows the user to specify fixed field, and variable fields which may have different data depending upon circumstances. Data can be transferred in blocks at a high speed, and either the full 3,000 character memory or individual sections of memory called "messages" can be dumped, onto cassette or through a computer to magnetic tape.

If further editing is required, the material on cassette can be played back into the Telterm2 memory and displayed on the screen. Hard copy is available by printing the cassette onto a hard copy terminal or listing the magnetic tape on a high speed printer. The Telterm2 can be configured to handle two cassettes, one for input, the other for output.

Table 6
TELTERM[13]

Item	Breakdown	Total
Telterm2 purchase		$6500
Telterm2 rental	$367.50	$4410.00 (12 mo. @ $367.50 ea.)
Cassette cost	$5 ea. (need 100 @ 50,000 ch. ea.)	$500 (100 @ $5 ea.)
Conversion to magnetic tape	$25/hr. of computer time	$250 (10 hrs. @ $25 ea.)
Total if leasing		$5160.00
Total if purchasing		$7250.00

ATS (Administrative Terminal System [Formerly IBM's DATATEXT])

The ATS is an on-line software system for storing information on disk and allowing direct editing through terminal-entered commands. Significant editing capability includes formatting ability to change number of characters per line, or number of lines per page. Typing

[13] Prices quoted by Mr. Larry Lohr, Data-Serv, 15114 Downey Avenue, Paramount, California 90723, (213) 531-6161. Price includes a Mobark 400T cassette attachment ($1800). It would be possible to use the Livermore cassette attachments ($375) with slight inconvenience. Lease is a lease-purchase on two year purchase commitment.

errors made and caught during entry are correctable merely by backspacing and retyping. Such changes to content as insertions, deletions, substitutions, and rearrangements of words, and phrases are made on-line by specifying the changes themselves and their locations. The user must specify enough unique information on an item to allow for exact scan matches, as well as specifying line numbers locations. ATS has playback capability (at any point in the editing process) of 140 words per minute on typewriter terminal, and 500 lines per minute on high speed printer. ATS (as implemented by Arcata Data Systems) also can output to microform and/or to a computer typesetting device automatically and directly.

In addition, the ATS system has the capability of producing, in interactive mode, the equivalent of one line KWIC's and counts of occurrences from which frequencies can be easily computed.

Material on disk can be dumped to magnetic tape at minimal cost when extended periods of hand-editing are required on hard copy. Such a procedure saves disk storage costs which are computed on a monthly basis.

Table 7
ATS[14]

Item	Breakdown	Total
Terminal purchase		$4700
Terminal rental	$100/mo. (requires Modem hookup - possible additional $25/mo.)	$1200-$1500 (12 mo. @ $100-$125/ea.)
Storage cost	$.15/1550 characters/mo.	$6000 (5,000 ch. for 12 mo.)
Conversion to magnetic tape	$10 (on or off disk)	$20
Total if leasing		$7220-$7520
Total if purchasing		$10,720

Optical Character Scanning (FormScan)

OCS is an on-line scanning system using as input a typescript on standard paper made with an IBM Selectric typewriter. Type fonts which can be read are 1403, OCR-A and Pica. Upper and lower case capability is available with Pica only, and a modified font must be obtained from FormScan to preclude O - Ø and l - 1 confusions.

[14] Prices quoted by Mr. Bruce Hawley, Butler Data Systems, 12911 Cerise Avenue, Hawthorne, California 90250, (213) 772-2331. Storage figure is shown in parallel form to cartridge and cassette cost, but is totally unrealistic in terms of actual use of system, since in real use, the majority of data would be kept on magnetic tape, saving storage costs, and only dumped onto disk for short periods for editing. A more realistic figure would be closer to $3000, giving a total of $4210 to $4510.

The non-reflecting areas (i.e., the black letters) are read and coded by a CRT and transferred to the character identifying element of the system. Character recognition is accomplished through the division of each 1/10" X 1/10" character area into a grid with 1200 parts and software identification of the grid areas which contain data specific to a particular character (ex. a or j or @). Identification error is at a maximum of 1 in 25,000 characters for upper and lower case and much lower for all caps. The system produces 7 track, 556 BPI, BCD coded tapes directly and a conversion to 9 track, 800 BPI, EBCDIC coded tape is available for $15. Also available is an updating feature allowing lines with errors to be corrected by typing a sheet identifying the line number and then typing the line as it should be.

Table 8
OCS[15]

Item	Breakdown	Total
Terminal rental		
Storage cost		
Convention Typescript	.01/line	$666.67 (66,667) lines of 15 words ea. @ .01 ea.)
Conversion from MTST Typescript	3.20/cartridge approximately	$666.67
Update	$50.00 minimum	
Total of leasing		$666.67
Total of purchasing		Not applicable

[15] Prices quoted by Mr. Harley D. Hancock, FormScan Inc., 16220 Orange Avenue, Paramount, California 90723, (213) 636-2441. Costs are deceptive since they do not reflect cost of preparation of a perfect typescript to be scanned.

Table 9 provides a quick comparison of the various input systems. In cases where some systems lack capabilities found in others, information in addition to the relative capability may be helpful. The following comments relate to each system which lacks a specific capability found in other systems.

1. Upper and lower case

   OCS - upper and lower case available only with pica

   CRT - upper and lower case is optional on Telterm2

2. Produces ASCII coded magnetic tape

   MTST - produces EBCDIC

   ATS - produces EBCDIC

   OCS - produces BCD

3. Unlimited insertion

   KEYPUNCH - insertion limited by number of blanks left at end of card

   TELETYPE - insertion practically impossible

   OCS - insertion practically impossible

4. Update facility

   KEYPUNCH - update requires card duplication and is limited by number of blanks left at end of card

5. Off-line keying

   ATS - it is possible to load a file created off-line onto the disc

6. Off-line edit

TELETYPE - computer edits using "control" characters

DATAPLEX - computer edits using correction cassette

ATS - computer edits using terminal input commands

7. Backspace correct

KEYPUNCH - requires duplicating whole card

TELETYPE - uses "control" characters

FLEXOWRITER - uses "control" characters

8. Needs no line feed with carriage return

TELETYPE - needs line feed

CRT - needs line feed

9. Initial playback of hard copy

KEYPUNCH - requires computer listing of cards

DATAPLEX - none available

CRT - cassette can be listed on computer printer or terminal by phone line

10. Hard copy of edited file

KEYPUNCH - requires computer listing of cards

TELETYPE - paper tape must be carried to processor

11. No hand carrying to magnetic tape

KEYPUNCH - card deck must be taken to computer location

MTST - cartridges must be taken to converter location

12. Untrained operator

KEYPUNCH - six weeks training

MTST - three days provided free by IBM

13. Uses other than input preparation

TELETYPE - can be used as terminal

MTST - can be used for secretarial tasks

CRT - can be used as terminal

Table 9

COMPARATIVE FEATURES OF THE INPUT DEVISES

CAPABILITY	Keypunch	Teletype	Flexowriter	MTST	Dataplex	CRT	ATS	OCS
1. Upper & lower case			X	X	X	X	X	X
2. Produces ASCII coded magnetic tape	X	X	X		X	X		
3. Unlimited insertion				X	X	X	X	
4. Update facility				X	X	X	X	X
5. Off-line keying	X	X	X	X	X	X		X
6. Off-line edit	X	X	X	X		X		X
7. Backspace correct				X	X	X	X	
8. Needs no line feed with carriage return			X	X	X		X	
9. Initial playback of hard copy		X	X	X			X	
10. Hard copy of edited file		X	X	X	X		X	X
11. No hand carrying to magnetic tape		X	X		X	X	X	X
12. Untrained operator		X	X		X	X	X	X
13. Uses other than input preparation		X		X		X		

III. Test On 300 Word Sample

From the point of view of entering and editing input text, one of the most difficult problems is inserting identifying codes within the text to discriminate among categories of information relevant to the analysis being done. For example, a linguist might wish to mark each relative clause or noun phrase with a code so that he could obtain information on their use as well as information on single words. Because the complexity of inserting such codes into a running text represents one of the most difficult problems for any input system to handle, it was used as a task to test the systems under study.

A 300 word sample representing a hypothetical transcription of audio-recordings was used for a test. There were copies of two transcripts in three stages of editing: (1) first typescript from audio tape; (2) intermediate hand-coded rough; (3) final copy. On stage 1, examples of simple typing errors (e.g., misspelled words) were inserted for immediate backspace correction. On stage 2, examples of types of codings or changes which might occur were inserted. Codings used were arbitrary and have no actual significance. As many types of editing problems as possible were incorporated within the example.

Example of Sample Used

Stage 1:	%AH	THIS	GUY	I.	HE	MGTH
	HIS,	HS	INTENTINS	MIGHT	BE	GOOD,
	YOU	KNW,	BUT	HE'S	JUST	DOING
	HIS	JOB,	YOU	KNOW?	RIGH?	(T7558)
	%AE	TRUE,	I	AGRE	WITH	YOU

Stage 2:

%AH	THIS	GUY,	I	HE	MIGHT,
HIS,	HIS,	INTENTIONS	MIGHT	BE	GOOD,
YOU	KNOW,	BUT	HE'S	JUST	DOING
HIS	JOB,	YOU	KNOW?	RIGHT?	(T7758)
%AE	TRUE,	I	AGREE	WITH	YOU

Stage 3:

```
$15.051 %Z1 THIS 12011 GUY, 57420
I 72100 HE MIGHT, HIS 72108
INTENTIONS MIGHT BE GOOD, 57300 57420
YOU 72208 KNOW, (BUT +) HE'S 57311
JUST DOING HIS JOB, YOU 72108 KNOW? 57441
RIGHT? (T7758)
%AM TRUE, I AGREE 5200 WITH YOU $$15.051
```

After arranging with the vendors for a test, they were asked to prepare the data, using trained operators and demonstrate each step from rough copy to magnetic tape. Following are the results of those tests:

KEYPUNCH

Tab setting was done by means of a program drum card. Set up for this card took approximately three minutes. Input typing proceeded at a rate of approximately 30 words per minute with an excessive noise level. During initial keying two errors were caught and corrected by duplicating the card up to the point of the error, correcting the error and duplicating the rest of the card. When the original test deck was finished, the operator attempted to perform the editing functions of adding codes, and changing items indicated. After only a few cards, it became apparent that almost every card was being retyped in order to make the corrections, and the test was abandoned.

TELETYPE

No tab setting was available. Input typing proceeded at a rate of approximately 25 words per minute with some confusion caused by the need to multiple space between words to simulate tabs. Noise level was high. Two control character corrections were used on initial keying, and proofreading indicated one additional keying error to be corrected during editing. There was no take up reel for paper tapes produced. Editing proceeded by playing back the original paper tape and attempting to stop it at appropriate points for insertions. After approximately fifteen minutes and dozens of attempts, the operator admitted she was unable to perform even the first insertion and the test was abandoned.

FLEXOWRITER

Tab setting for initial input was smooth and uncomplicated. Input typing proceeded at a rate of approximately 30 words per minute with a high but acceptable noise level. There was no take up reel for paper tape produced.

Editing proceeded by playing back the original tape and stopping at appropriate points for insertions. A new paper tape was punched during editing. Stopping was accomplished at the proper point in all but one instance. The entire new tape had to be run through another pass of the editing procedure for the final correction to be made.

MTST

Tab setting for initial input was smooth and apparently uncomplicated. Input typing proceeded at a rate of approximately 40 words per minute

with an acceptable noise level. The process of setting up MTST tapes for editing (rewinding, moving tape to other side of machine, etc.) took approximately one minute.

Editing was accomplished by playing back the original tape, stopping at appropriate points and typing in the insertion. No special commands were required. A single button accomplished playback and only the actual characters to be inserted were keyed in the editing stage.

As many as 130 characters may be keyed per record initially and there is automatic creation of a new record if editing insertions cause an overflow. There is no automatic facility for removing superfluous blanks.

DATAPLEX

Although the specifications of the task had been specified in advance, Data Instruments was unable to have a system ready to view. The explanation given was that computer software was required specific to the task in order to have their machines process material. The user was expected to provide programmer time to write such software which had to be written in Data Instruments own programming language. Assurances were given that once such software existed, the system could handle any contingency. There was, of course, the understanding that the efficiency of the system would be directly dependent on the sophistication of the software the user produced.

CRT

Tab setting for initial input was smooth and uncomplicated. Input typing proceeded at approximately 30 words per minute with virtually no

noise. "End of message" codes were keyed after 1500 characters (20 lines). When the 3000 characters buffer was filled, the operator dumped the buffer, one message at a time, onto the cassette attachment. Setting up the tape took two simple operations. The terminal was switched from LOCAL to ON-LINE for transfer, and from TYPE to TELETYPE mode. There were then four transfer commands entered on the keyboard.

Transfer proceeded smoothly. Hard copy was obtained by playing back the cassette through a teletype. Setting the tape took two simple operations.

Data was then transferred from the cassette back onto the Telterm2 for further editing. Insertion was accomplished by moving the cursor to the appropriate point, pushing the "Insert start" key and typing in the insert. Editing proceeded swiftly and surely. Superfluous blanks were removed with the "delete character" key. When editing was completed, the operator again simulated transfer to cassette.

ATS

Tab setting for initial input had to be done twice due to operator error. Input typing proceeded at a rate of approximately 40 words per minute with an acceptable noise level. There was an editing set-up procedure requiring two commands, taking only a few seconds.

The editing procedure proved to be unacceptable, due to the fact that the commands required by the system necessitate the typing of the full word occurring before the desired insertion. In this particular application, with its initial one-to-one expansion, it would be more efficient to retype.

Several other features of interest were demonstrated, however. ATS has the capability of scanning for a particular character set (e.g., a specific linguistic code) and giving immediately the number of occurrences. It can also print out each line on which such a character set occurs (roughly equivalent to a one line KWIC). For the purposes of initial "browsing" in data, ATS may prove to be highly cost-effective.

OCS

Since the preparation and editing of text is not a function of scanning, no parallel test was made of the OCS system. The operation of the scanning equipment was reviewed for accuracy and ease of conversion.

A stack of typed sheets was input to a hopper which fed them into the scanning section of the machine. Transfer to magnetic tape proceeded smoothly until a character was encountered which the machine could not recognize. This character (a smudged "m") was projected on a screen and the operator entered the correct choice through a teletype. There were no other errors. The magnetic tape was then played back through a teletype for final proofreading and hard copy.

IV. Conclusions

After all systems were evaluated, it became clear that two, MTST and Telterm, were definitely superior for the task of preparing a large natural language data base which required heavy editing and insertion.

In each of the other systems reviewed, correction of typing errors caught during entry was either cumbersome, time-consuming, or virtually impossible without retyping a large section of the data. If the data was keypunched, a card with an error required repunching. In the case of the Teletype, such corrections required an on-line edit performed by a computer. ATS used both a computer and a cumbersome line referencing and retyping process. In the cases of MTST and Telterm, such corrections could be made directly, immediately, and simply.

In most of the other systems, large scale insertions were impossible or extremely difficult. On Telterm, such insertions could be accomplished by simply pushing the "insert" button and typing in the new data. On MTST, insertions could be made by playing the data back to the point at which the insertion was to begin, typing in the new data, and then playing back the old data up to the next insertion.

MTST have the advantage over, TELETYPE, FLEXOWRITER, DATAPLEX, and ATS in that they can do <u>all</u> the required data preparation and editing off-line, thus requiring no costly computer time.

Both MTST and Telterm are significantly better than other systems. Both have equal flexibility and editing capabilities for text input. Any final decision on which of the two is the most effective method for a particular project requiring text preparation must be based on a number of factors not directly examined in this study, but relating to the project itself, its task, funding and scope.

If the particular project is one which requires hard copy at every step, MTST would be preferable, as it would be if the researcher is in

a non-computer oriented environment, where he would have no use for the terminal capabilities of the Telterm after completion of data preparation, but could use the MTST in normal secretarial routines.

If, on the other hand, the project is in a computer oriented environment where the terminal capabilities would insure continued use for the terminal, Telterm would be the logical choice. Likewise, all other factors being equal, Telterm would be the preferred device, since MTST costs $2,875 per year more than Telterm on a one year lease in basis, and $7,010 more on a purchase basis.

Above all else, one fact emerges clearly from this study. The rapid proliferation of sophisticated input techniques has assured the researcher natural language tools which eliminate previous roadblocks and allow him to concentrate his energies to a greater extent on the research itself, rather than the mechanics of data preparation.